A Research Proposal

on

Recognition of Degraded Devanagari Text

Submitted to

LOVELY PROFESSIONAL UNIVERSITY

in partial fulfillment of the requirement for the award of degree of

DOCTOR OF PHILOSOPHY (Ph.D.) IN (Computer Science and Engineering)

Submitted by:

Rajni Kumari

Supervised by:

Dr. Gurpreet Singh Lehl

FACULTY OF TECHNOLOGY AND SCIENCES
LOVELY PROFESSIONAL UNIVERSITY
PUNJAB

Introduction

In the present digital world, it has become mandatory to have all the available information in a digital form recognized by machines. In country like India, where there is abundance of information in the form of manuscripts, ancient texts, books, etc. that are traditionally available in printed and handwritten form, such printed material are in-adequate when it comes to searching information among thousands of pages. It has to be digitized and converted to a textual form in-order to recognize by machines doing searches of a million pages / second. We also have to retrieve and extract the text which is deteriorated. Then only, the true knowledge of Indian history, tradition and culture would be available to the masses and the digital revolution would be said to have reached the information age.

In the field of pattern recognition and artificial intelligence Optical Character Recognition (OCR) has become one of the most successful applications. Today, reasonably efficient and inexpensive OCR packages are commercially available to recognize printed texts in widely used languages such as English, Chinese, and Japanese. OCR is the most essential part of Document Analysis System that converts the scanned images of text , books, magazines, and newspapers into machine-readable text. The process of Document Analysis Recognition can be divided into two parts, namely, printed and handwriting character recognition. The printed documents can further be divided into two parts: good quality printed documents and degraded printed documents. Degradation of the text leads to touching, broken, heavy printed, typewritten, faxed document, back side visible, manually underlined lines of text and digital image from camera figure 1.1.

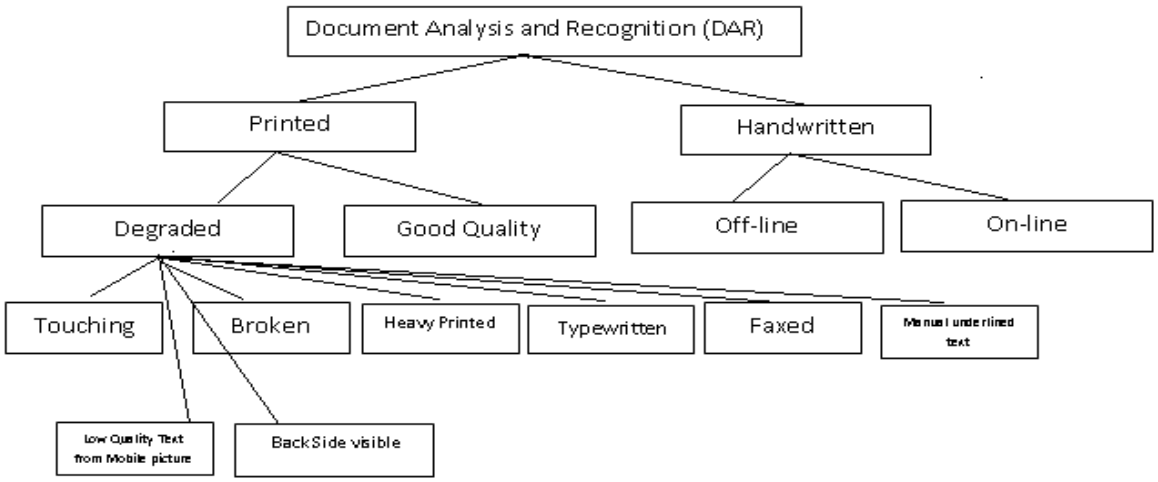

Figure: 1.1
Categories of DAR

Recognition of degraded documents has been an active research area in the field of pattern recognition. Handwriting character recognition has been divided into offline and online character recognition, as shown in figure 1.1. Offline documents are scanned images of prewritten text, generally on a sheet of paper. In online handwriting recognition, data are captured during the writing process with the help of a special pen and an electronic surface. All Indian scripts are cursive in nature which makes them hard to recognize by machines. Scripts like Devanagari, Gujarati, Bengali and many others have conjuncts or joint-characters increasing segmentation difficulties. To add to that, various fonts of various sizes used for printing texts over the years, the quality of paper, scanning resolution, images in texts and degraded quality of text and paper etc puts challenges for the researchers for image processing job. Also, it requires huge linguistic know-how to apply post-processing.

Devanagari is the script used for writing many official languages in India such as Hindi, Marathi, Sindhi, Nepali, Sanskrit and Konkani, where Hindi is the national language of the country. Hindi is also the third most popular language in the world [10]. Several other Indian languages like Gujarati, Punjabi, and Bengali use scripts similar to Devanagari. More than 500 million people use Devanagari script for documentation in central and northern parts of India [11].

There are large collections of data originating in paper form, such as office data, books, newspapers, and historical manuscripts written in Devanagari but some of them are highly degraded (figure1.1), that would be benefitted by being available in digital form. Paper documents are ubiquitous in our society, but there is still a need to have many of these documents in electronic form. Once these documents are converted into electronic form, the documents can be used in digital libraries or for wider dissemination on the internet. To increase efficiency of paperwork processing and to better utilize the content of data originating in paper form, it is required that the paper documents be converted into images of the documents in electronic form, and that must be converted to a computer-readable format such as ASCII format. This will allow editing as well as search and retrieval of information on these documents. The conversion is done by scanning the document and then processing the digital image with an OCR algorithm. A typical OCR consists of sequence of activities involving Digitization, Preprocessing, Segmentation, Features Extraction, Classification and Post Processing. Digitization is the process of converting the paper based degraded Devanagari document into electronic form through imaging a process whereby a document is scanned and an electronic representation of the original, in the form of a bitmap image, is produced. To convert the digital image to simple ASCII strings, a crucial step is to extract only the pixels that belong to the characters to be recognized. Digitization produces the digital image, which is feed to the pre-processing phase. Preprocessing is the initial stage of character recognition and is used for skew detection/correction, skeletonization, and noise reduction/removal. Noise removal is used to remove unwanted bit pattern which does not play any significant role in document. Skewness refers to the tilt in the bit mapped image of the scanned paper for OCR. It is usually caused if the paper is not fed straight into the scanner. Skeletonization is used for decreasing the line width of text from many pixels to single pixel. Thinning is the process of peeing off a pattern as many as pixels as possible without affecting the general shape of the pattern. Noise removal is used to remove unwanted bit pattern which does not play any significant role in document After preprocessing, digital document that goes to segmentation phase that extracts lines, words and then finally characters from the text document images if required. This phase is very necessary for recognition. For the task of segmentation, an algorithm is use to find the segmentation point in the text. The challenge of a segmentation technique lies in the decision of best segmentation point for line, word and character isolation. The purpose of feature extraction is

to find a subset of features that will maximize the effectiveness of recognition, or maximize the efficiency of the process with or without the involvement of the classifier. This phase is used to measure the relevant shape contained in the character. In the feature extraction phase, one can extract the features according to levels of text, e.g., character level, word level, line level and paragraph level. Features can be also selected from documents with different sources like font, style and stroke detection etc. The goal of feature extraction is to map input patterns onto points in a feature space, the purpose of classification is to assign each point in the space with a class label or membership to the defined classes. It is the decision making phase of an OCR engine. The preliminary aim of classification phase of OCR is to develop the constraint for reducing the misclassification relevant to feature extractions. Classification is very difficult task for unknown patterns. OCR results usually contain errors because of character classification and segmentation problems. For the correction of recognition errors, OCR systems apply contextual post-processing techniques. The two most common post-processing techniques for error correction are dictionary lookup and statistical approach. The advantage of statistical approach over dictionary-based methods is computational time and memory utilization. Conversely, lexical knowledge about entire words is more accurate when using a dictionary.

Recognition of degraded Devanagari text means that machine should be able to recognize the text which is highly broken, touched and noisy. Recognition has the challenges of language because of its complexity and loss of structural information because of degradation. Devanagari is an alphabetic script with a complex composition of the constituent symbols. Devanagari script has 13 vowels (Fig. 1.2.(a)) and 34 consonants (Fig. 1.2.(b)) along with 14 modifiers (11 of vowels and 3 of ‗rakar' as shown in Fig1.2(a)) symbols. Apart from vowel and consonant characters, called basic characters, there are compound (composite) characters in most of Indian scripts including Devanagari, which are formed by combining two or more basic characters.

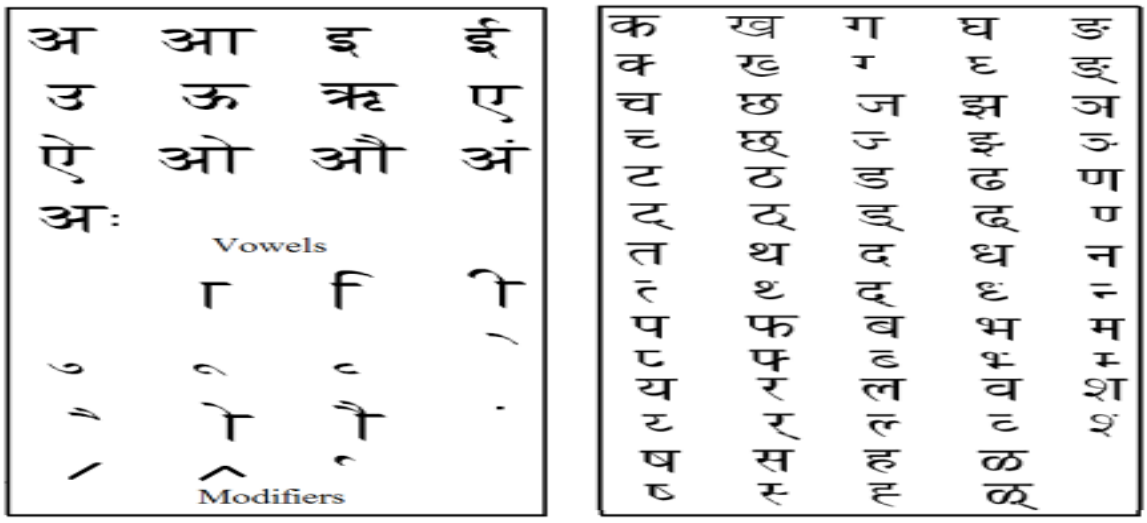

Figure 1.2
(a) The vowels and modifiers of Devanagari script
(b) Consonants and their corresponding half forms (shown below the consonants) in Devanagari script

The shape of a compound (composite) character is usually more complex than the constituent basic characters. A vowel following a consonant may take a modified shape, which depending on the vowel is placed to the left, right, top or bottom of the consonant; they are called modifiers or _matras'. Another distinctive feature of Devanagari script is the presence of a horizontal line on the top of all characters. This line is known as header line or shirorekha (See Fig.1.3). Regular devanagari words can typically be divided into three strips: top, core or middle and bottom as shown in Figure 1.3. The header line (shirorekha) separates the top strip and the core strip and base line separates core strip and lower strip. The top strip generally contains the top modifiers, and bottom strip contains lower modifiers.

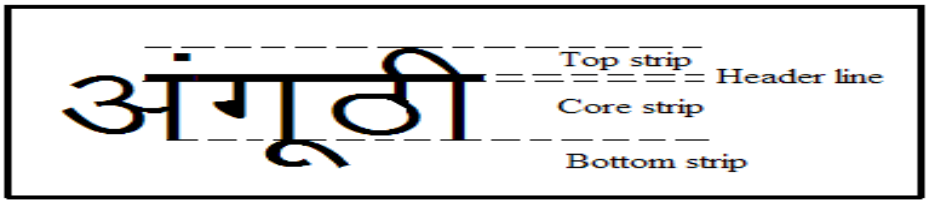

Figure 1.3.

Three strips of a word in Devanagari script.

Every Indian script has its own specified composition rules for combining vowels, consonants and modifiers. Some of them can be combined with their

type as well as shown in figure 1.4 modifier can be attached to a vowel or to a consonant.

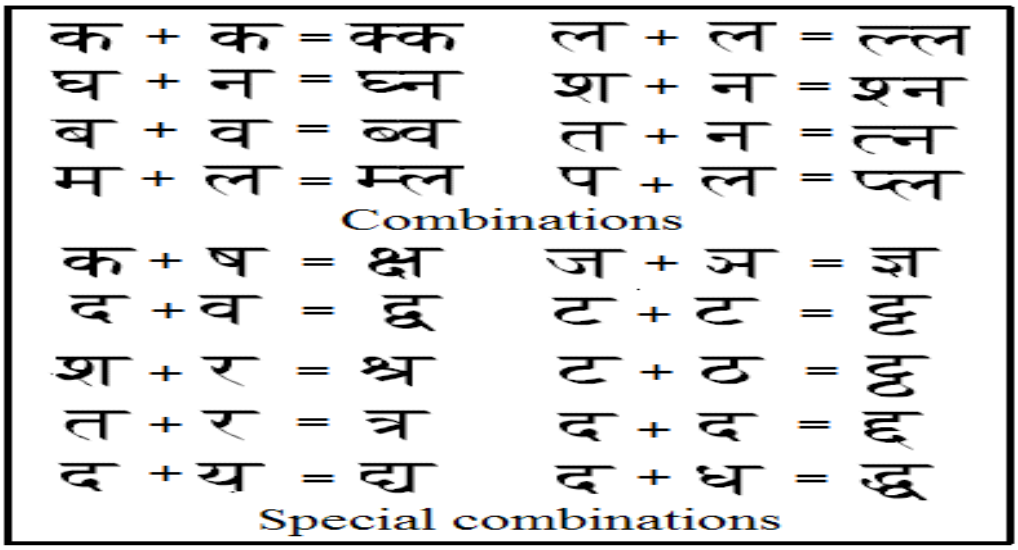

Figure 1.4

Some combinations of consonants with themselves.

Literature Survey

A lot of work has been reported in the recognition of Indian scripts. The pioneer being the *R.M.K Sinha [47, 48]* who in 1970 initiated automatic recognition of printed Devanagari script at Indian Institute of Technology Kanpur. He presented a syntactic pattern analysis system to Devanagari script recognition [13]. Palit and Chaudhuri [14] as well as Pal and Chaudhuri [15] also has presented another OCR system for printed Devnagari. A team comprising Prof. B. B. Chaudhuri, U. Pal, M. Mitra and U. Garain of Indian Statistical Institute, Kolkata, developed the first commercial level product for printed Devanagari OCR. In Satish [5] some common irregularities in writing Devanagari script are reported here. Jindal and Lehal [6-9] reported work on degraded Gurumukhi text. R.M.K. Sinha and H.N. Mahabala [2] once tried to recognize Devanagari automatically according to their pattern analysis system. They choose 26 symbols and extracted structural information for these characters. However, their experiment was limited in sample size, and could not give quantitative recognition rate. A significant work involving preprocessing, feature extraction,

classification and post-processing techniques has also been reported so far. An approach based on the detection of shiroreka is proposed by Chaudhuri and Pal [16] with the assumption that the skew of such header lines show the skew of the whole document. The authors addressed the problem of a small degree of skew (tilt) affected the header line (shirorekha) joining all the characters in a word. There are some documents called multi-skew documents in which the text lines are not parallel to each other. In [17], an approach, which is an extension to that of [16] is proposed for skew estimation in multi-skew documents. Das and Chanda in [18] also proposed a fast and script independent skew estimation technique based on mathematical morphology. In segmentation phase after skew elimination, the separation of paragraphs, text lines, words and characters is to be carried out for effective feature extraction. In *Projection profiles*, space between words and lines are used to achieve this in [12]. In [46, 54] the character segmentation is done by Removal of shirorekha (header line). Govindaraju et al. [19] distinguished text pattern from graphics (images) by analyzing the shape of the horizontal projection profile and the regularity of intervals between the peaks in the profile and their width. Garain and Chaudhuri [20] presented the technique based on fuzzy multi-factorial analysis for identification and segmentation of touching machine printed Devanagari characters. An algorithm was developed for effectively selecting possible cut columns for segmenting the touching characters. Bansal and Sinha [21] presented a two-pass algorithm for the segmentation of machine printed composite characters into their constituent symbols. The proposed algorithm extensively uses structural properties of the script. In the first pass, words are segmented into easily separable characters or composite characters. Statistical information about the height and width of each separated box is used to hypothesize whether a character box is composite. In the second pass, the hypothesized composite characters are further segmented. Kompalli et al. [23] used a graph representation method to segment characters from printed words. A classifier then processes the components and the classifier score is used to determine if the components need to be further segmented. Sharma et al. [25] presented a rule-based approach for skew correction along with removing insignificant data like dark band, thumb mark and specks. The proposed system is based on rule-based methods, morphological operations and connected component labeling. In the method proposed by Kompalli et al.[4], the shirorekha is determined using projection profile and run length. Structures above the shirorekha are isolated as ascenders. Average character height is used to predict a region of interest (ROI) to

locate descenders. Once the shirorekha is removed, for Devanagari, the top, middle and bottom zones are identified easily. Components in top and bottom zones for Hindi (Devanagari script) are part of vowel modifiers. Each of these components is then scaled to a standard size before feature extraction and classification Jawahar et al. [26]. For the binarization of natural scene images containing Devanagari textual information, an adaptive thresholding technique is proposed in [28], which uses the simple average gray value in a window of size 27 × 27 around a pixel as the threshold value for that pixel. A water reservoir based analogy is proposed in U.pal et al [29] to extract individual text lines from such documents. In S.Sinha et al [34-35] for word-wise script identification, the document is initially segmented into lines and then the lines are segmented into words. Individual script words are identified from document images using different topological and structural features including number of loops, header line feature, and water-reservoir concept based features, profile features, etc. Different features have been used for the recognition of Devanagari characters. The system described by Sinha and Mahabala [14] for printed Devanagari characters stored structural descriptions for each symbol of the script in terms of primitives and their relationships. Sinha [36] also demonstrated how the spatial association among the constituent symbols of Devanagari script plays an important role in understanding Devanagari words. In V.Bansal et. al[1] a character is assigned to one of the three groups namely basic, modifier and compound character groups and group-wise features are considered. Also it is observed that the compound characters (around 250) in the script occupy only 6% of the text. The basic, modifiers and compound character sub grouping is done using the following properties: (i) bounding box width, (ii) number of border pixels per unit width, and (iii) accumulated curvature over the border per unit width. V.Bansal et. al[1] considered several statistical classifying features like horizontal zero-crossings, moments, vertex points and pixel density in different zones for Devanagari characters. Also they considered word envelop information containing number of character boxes, number of vertical bars, number of upper modifier boxes, number of lower modifier boxes, vector giving position of vertical bars, vector giving type and position of each character box. The major two features considered for printed Devanagari characters by Jayanthi et al. [37] are main horizontal line and various vertical lines. The third feature is to test whether vertical lines are present in the rightmost side of the character. The other features have been the height to width (aspect) ratio of the character, whether the character is narrow or broad ended

and the number of free ends it has. Govindaraju et al. [38] considered gradient features for feature selection of the characters. The gradient features computed using a Sobel operator measures the magnitude and direction of intensity changes in a small neighborhood of each pixel. A gradient map is computed by retaining those gradient magnitudes, which are above a threshold. The feature vector is constructed by concatenating the gradient vectors of the constituent blocks. Kompalli et al. [39] used gradient, structural and concavity (GSC) features for OCR of machine printed, multi-font Devanagari text. The gradient features were used to classify segmented images. In the method proposed by Dhurandhar et al., [40], the significant contours of the printed character are extracted and characterized as a contour set based on a reference co-ordinate system. Jawahar et al. [26] used principal component analysis (PCA) for feature extraction of printed characters. A word-level matching scheme for searching in printed document images is proposed by Meshesha et al. [41]. The feature extraction scheme extracts local features by scanning vertical strips of the word image and combines them automatically based on their discriminatory potential. The features considered are word profiles, moments and transform domain representations. In [3], printed Hindi words are initially identified from bilingual or multilingual documents based on features of the Devanagari script using SVM. Identified words are then segmented into individual characters in the next step, where the composite characters are identified and further segmented based on the structural properties of the script and statistical information. In [31], a technique to identify Kannada, Hindi and English text lines from a printed document was presented. The system is based on the top-profile and bottom-profile of individual text lines of the input document image. The positions of the connected components of top and bottom profiles are extracted and the coefficients of variation of the top and bottom profiles are computed thereafter. The features used for script identification of machine printed text in [30] are 64- dimensional chain-code-histogram features and 400-dimensional gradient features. For the purpose of indexing in [42], printed Devanagari word images are represented in the form of geometric feature graphs (GFG). It is a graph-based representation of the features extracted from the image of the word. A set of features including percentiles, horizontal and vertical derivatives of percentiles, angles, correlations, and energy were used for the recognition of printed Devanagari character recognition in [43]. Linear discriminant analysis was then used to reduce the dimensionality of the feature set from 81 to 15. Zernike moments and directional features are used as the features for printed

characters in [44]. There are printed complex documents where text lines of a single page may have different orientations or the text lines may be curved in shape. As a result, it is difficult to detect the skew of such documents and hence character segmentation as well as recognition of such documents is a complex task. Using background and foreground information a scheme towards the recognition of Indian complex documents is proposed in [45]. Convex hull and water reservoir principle based features as well as different angular information obtained from the external and internal contour pixels of the characters have been applied for the recognition of multisized/ multi-oriented characters in Devanagari and Bangla text [46-50].

OCR classification techniques such as Template Matching, Statistical Techniques, Neural Networks, Support Vector Machine (SVM) algorithms, Combination classifier have been used for Devanagari character recognition[51-55]. In the training phase of such techniques, patterns are fed and the system adjusts itself to minimize the misclassification errors of these patterns. Such trained systems are used for the classification of unknown test patterns. One finest example of such a system is the artificial neural network, which adjusts the weights of its links from the training patterns for better classification. Neural network architectures can be classified as, feed-forward and feedback (recurrent) networks[56-59]. The most common neural networks used in the OCR systems are the multilayer perceptron (MLP) of the feed forward networks and the Kohonen's Self Organizing Map (SOM) of the feedback networks. MLP is proposed by U. Bhattacharya et al. [60]. A detailed comparison of various NN classifiers is made by M. Egmont-Petersen [61]. He has shown that Feed-forward, perceptron higher order network. K. Y. Rajput et al. [62] used back propagation type NN classifier. The support vector machine (SVM) is a new type of pattern classifier based on statistical learning techniques. SVM was originally defined for two-class problem and it looks for the optimal hyper-planes that maximize the distance, the margin, between two classes. Shailedra et al. has implemented binary classification technique of Support Vector Machine (SVM) using linear kernel function which produces 99.48% overall recognition rate. Jawahar et al., [26] presents an SVM-PCA based OCR for two major Indian languages which attains an overall accuracy of approximately 96.7%. Extensive experimentation is carried out on an independent test set of approximately 200000 characters. HMM-based systems provide a language-independent framework for training and recognition. At the same time, they do not require the training data to be segmented into words or characters, i.e., they automatically

train themselves on non-segmented data. Dhingra et al., [48] had used Hidden Markov Model (HMM) for degraded documents, such as faxed text to develop robust OCR. As indicated by experimental results, a properly chosen training set can significantly reduce the recognition error rate. A top-down binary tree based recognition of printed Devanagari characters was proposed by Jayanthi et al. [37] as binary tree is one of the fastest decision making processes for a computer program. Govindaraju et al. [38] considered 38 characters and 83 frequently occurring conjunct character classes in a multistage classification approach. Initially they were classified into four categories depending on their structural properties. The recognition of printed characters in the method proposed by Dhurandhar et al. [40] involved comparing the contour sets with those in the enrolled database. Subsequently, classification using K-nearest neighbor classifier is performed with the assumption of accurate segmentation.

Problem Definition:

Recognition of degraded Devanagari text means that machine should be able to recognize the text which is highly broken, touched and noisy. Recognition has the challenges of language because of its complexity and loss of structural information because of degradation (Fig1.5). When two or more characters sit side by side to form a word in Devanagari, the header lines touch and generate a bigger header line. Since characters in a word touch in Devanagari, character segmentation from word is a difficult task in Devanagari OCR development. After scanning from poor quality text, highly broken characters are found which have lost its structural properties as well as its identification as shown in Figure 1.6.

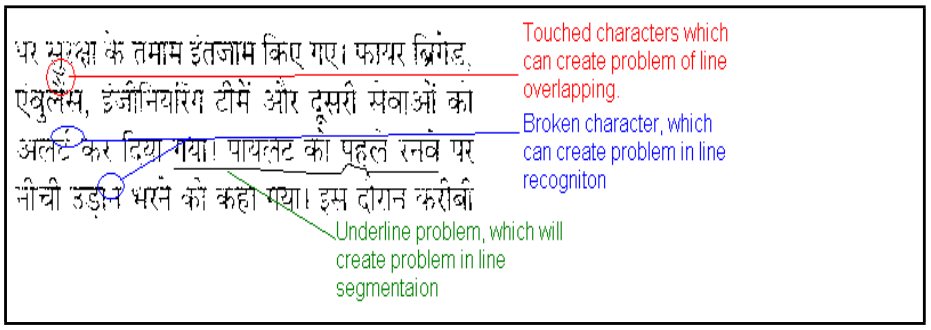
Touched characters which can create problem of line overlapping.
Broken character, which can create problem in line recogniton
Underline problem, which will create problem in line segmentaion

Figure 1.5.
Problems of Degradation

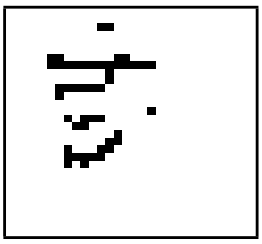

Figure 1.6
Highly degraded text of Devanagari

There is the loss of upper or lower modifier completely also in the text, which becomes really a challenge for recognition process. The problem of degradation of Devanagari text has been identified and handled by Bansal [24] and many other researchers, but highly degradation was not considered. Hence this is the area where lot of research is required, and our research is focusing on this area. The proposed work in this report addresses the problem of Recognition of Degraded Devanagari Text by developing different algorithms.

SIGNIFICANCE

There are many OCR systems available for handling the clean printed documents in Devanagari, Gurumukhi, English, Japanese, and Oriya, etc. But till now there is no software available for recognition of printed degraded Devanagari script. As such, there is a need for degraded OCR for Devanagari script that can help the people for converting the text to computer process able format. The software developed in the present work can be used for recognition of clean, touched documents of Devanagari script. There is a need to develop OCR, which can handle broken characters and half characters.

SCOPE OF THE STUDY

- Digitized data will be available for the masses for search.
- It will do the value addition to get the true knowledge of Indian history, tradition and culture.
- It will not be lost and degraded in near future.

Proposed Objectives and Methodology:

The broad outline of the proposed objective is as follows:

- To develop an algorithms, so that the OCR should be able to recognize highly degraded offline Devanagari text.

To achieve the above objectives, the following components of the OCR will be studied.

- To use some of the existing methods for preprocessing phase such as noise removal,/smoothing, document skew detection/correction, thinning, binarization, digitization.
- To explore existing methods of restoration for highly degraded document especially for the broken characters such as median filtering, weiner filtering, morphological operators , hybrid restoration techniques approach, markov model for restoration, joining broken characters using connected components, gabor filter and active contour model.
- To propose new algorithms for restoration of highly broken text.
- To explore existing features (structural and statistical) and to propose new features for degraded printed Devanagari text.
- To use Support Vector Machine [SVM] classifiers.

Proposed Research Methodology.

Following will be the assumption to carry out the proposed research.

1. The scanned documents will be in Binarization (digitization) form.
2. Single column printed text will be considered.
3. Text and non-text separation will be primarily defined.
4. Proposed new algorithms will be tested on 500 data. It will be scanned from old books, newspapers and magazines.
5. Textual Degradation (Broken characters , Touching characters: Line touching overlapping and Inter -Character touching , Bleed through text) and Non-textual degradation (Underline or other lines marks on the document and Back side

visible)

With the above assumptions, the work will further be carried out as described below:

- The proposed method will be able to restore the highly broken characters. The restoration will be done in two phases one before segmentation and second after character segmentations. The proposed method will use hybrid approach to handle broken characters and other degradation. Some of the existing methods such as gabor filter and active contour model along with devanagari structural properties will be used.
- The proposed method will solve the line overlapping segmentation problem. This method makes use of the projection profiles and CC with some heuristics to segment the overlapped text documents in a robust way.
- The segmentation phase shall be implemented. Its aim is to separate lines, word and character of degraded printed devanagari text. In this phase, we will review the Run-Length Smoothing Algorithm[RLSA], projection profile, connected components and contour analysis. New algorithm will be designed for segmenting the lines, words and characters from degraded printed devanagari text using structural features of the language.
- In feature extraction phase, we will review existing algorithms and techniques like foreground skeleton, background skeleton, stroke detection and elastic matching. We will also review the structural and statistical features for degraded printed devanagari text. Based on these the features shall be detected for degraded printed devanagari text.
- In classification phase SVM will be used.
- Post processing also will be implemented.

ACTIVITY SCHEDULE

Activity Schedule	**Time Scale**				
	Sep-2012	**Jan-2013**	**Feb-2013**	**March-2013**	**Sep-2013**
Literature survey					
Preprocessing					
Segmentation					
Feature extraction					
Classification					
Post –Processing & Communication of research papers					
Documentation, Thesis writing and submission					

REFERENCES

[1] V. Bansal, Integrating Knowledge Sources in Devanagari Text Recognition. Ph.D. thesis, Indian Institute of Technology, Kanpur, India, 1999.

[2] V. Bansal and R.M.K Sinha, Segmentation of touching and fused Devanagari characters, Pattern Recognition 35, 875–893., 2002.

[3] H. Ma, D. Doermann, ―Adaptive Hindi OCR using generalized Hausdorff image comparison‖, ACM Transactions on Asian Language Information Processing, Vol.2, No.3, pp.193-218. 2003.

[4] Kompalli, snayak, setlur govind, Challenges in OCR of Devanagari Documents, IEEE Proceedings of the 2005 Eight International Conference on Document Analysis and Recognition (ICDAR‘05) 1520-5263/05 ,2005

[5] Satish Kumar, An Analysis of Irregularities in Devanagari Script Writing – A Machine Recognition Perspective, International Journal on Computer Science and Engineering (IJCSE) Vol. 2, No. 2, 2010

[6] M. K. Jindal, G. S. Lehal and R. K. Sharma, ―On Segmentation of touching characters and overlapping lines in degraded printed Devanagari script‖, International Journal of Image and Graphics (IJIG), World Scientific Publishing Company, Vol. 9, No. 3, pp. 321-353, 2009.

[7] M. K. Jindal, R. K. Sharma and G. S. Lehal, ―Segmentation of Horizontally Overlapping Lines in Printed Indian Scripts‖, International Journal of Computational Intelligence Research (IJCIR), Research India Publications, Vol. 3, No. 4, pp. 277-286, 2007.

[8] M. K. Jindal, R. K. Sharma and G. S. Lehal, ―Segmentation of Touching Characters in Upper Zone in printed Devanagari Script‖, in Proceedings of 2nd Bangalore Annual Compute Conference on 2nd Bangalore Annual Compute Conference, (Bangalore, India, January 09 - 10, 2009). COMPUTE '09. ACM, New York, NY, 1-6.

[9] M. K. Jindal, R. K. Sharma and G. S. Lehal, –Structural Features for Recognizing Degraded Printed Devanagari Script‖, in Proceedings of the IEEE 5th International Conference on Information Technology: New Generations (ITNG 2008), pp. 668-673, April 2008.

[10] U. Pal, B. B. Chaudhuri, –Indian script character recognition: A Survey‖, Pattern Recognition, Vol.37 pp.1887-1899, 2004.

[11] R. M. K. Sinha, –A Journey from Indian Scripts Processing to Indian Language Processing‖, IEEE Annals of the History of Computing,pp.8-31, 2009.

[12] B. B Chaudhuri, U. Pal, –An OCR system to read two Indian language scripts: Bangla and Devanagari‖, In Proceedings of the 4th International Conference on Document Analysis and Recognition, , pp.1011–1015, 1997

[13] R.M.K. Sinha, –A Syntactic pattern analysis system and its application to Devnagari script recognition‖, Ph.D. Thesis, Electrical Engineering Department, Indian Institute of Technology, Kanpur, India, 1973.

[14] R M K Sinha, H Mahabala, –Machine recognition of Devnagari script‖, IEEE Transactions on Systems Man and Cybernetics, Vol.9, pp.435–441, 1979.

[15]. U. Pal, B.B. Chaudhuri, –Printed Devnagari script OCR system‖, Vivek, Vol.10, pp.12–24, 1997.

[16] B. B. Chaudhuri, U. Pal, ―Skew Angle Detection of Digitized Indian Script Documents‖, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol.19, No.2, pp.182-186, 1997.

[17] U Pal, M Mitra and B B Chaudhuri, –Multi-Skew Detection of Indian Script Documents‖, In Proceedings of the Sixth International Conference on Document Analysis and Recognition, pp.292-296, 2001

[18] A.K. Das and B. Chanda, ―A fast algorithm for skew detection of document images using morphology‖, International Journal on Document Analysis and Recognition, Vol.4, No.2, pp.109-114, 2001

[19] V. Govindaraju, S. Khedekar, S. Kompalli, F. Farooq, S. Setlur, R.Vemulapati, Tools for enabling digital access to multilingual indic documents‖, In Proceedings of the 1st International Workshop on Document Image Analysis for Libraries, pp.122–133, 2004.

[20] U. Garain and B. B. Chaudhuri, ―Segmentation of Touching Characters in Printed Devnagari and Bangla Scripts Using Fuzzy Multifactorial Analysis‖, IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol.32, No.4, pp.449-459, 2002.

[21] V. Bansal, R. M. K. Sinha, ―Segmentation of touching and fused Devanagari characters‖, Pattern Recognition, Vol.35, pp.875–893, 2002.

[23] S Kompalli, S Setlur and V Govindaraju, ―Devanagari OCR using a recognition driven segmentation framework and stochastic language models‖, International Journal on Document Analysis and Recognition, Vol.12, pp.123–138, 2009.

[24] V. Bansal, R. M. K. Sinha, ―Integrating knowledge sources in Devanagari text recognition‖ , IEEE Transactions on Systems Man and Cybernetics Part A: Systems & Humans , Vol.30, No.4, pp.500–505, 2000.

[25] P. K. Sharma, K. D. Dhingra, S. Sanyal, ―A Rule Based Approach for Skew Correction and Removal of Insignificant Data from Scanned Text Documents of Devanagari Script‖, in Proceedings of the 3rd IEEE Conference on Signal-Image Technologies and Internet-Based System, pp.899-903, 2008.

[26] C. V. Jawahar, P. Kumar, S. S. R. Kiran, ―Bilingual OCR for Hindi-Telugu Documents and its Applications‖, in Proceedings of the Seventh International Conference on Document Analysis and Recognition, pp. 1-5, 2003.

[27] U. Garain, B.B. Chaudhuri, "On OCR of Degraded Documents Using Fuzzy Multifactorial Analysis", in Proceedings of AFSS International Conference on Fuzzy Systems, pp.388-394, 2002.

[28] U. Bhattacharya, S. K. Parui, S. Mondal, "Devanagari and Bangla Text Extraction from Natural Scene Images", in Proceedings of 10th ICDAR, pp.171-175, 2009

[29] U. Pal, P. P. Roy, ―Multi-oriented and curved text lines extraction from Indian documents‖, IEEE Trans. on Systems, Man and Cybernetics- Part B, Vol. 34, pp. 1676-1684, 2004.

[30] S. Chanda, S. Pal, K. Franke, U. Pal, "Two-stage Approach for Word-wise Script Identification", in Proceedings of 10th ICDAR, pp.926-930, 2009.

[31] P. A. Vijaya, M. C. Padma, "Text Line Identification from a Multilingual Document", in Proceedings of International Conference on Digital Image Processing, pp.302-305, 2009

[32] P.S. Hiremath, J.D. Pujari, S. Shivashankar, V. Mouneswara, "Script identification in a handwritten document image using texture features", in Proceedings of International Advance Computing Conference, pp.110-114, 2010.

[33] T. N. Vikram, D. S. Guru, "Appearance based models in document script identification", in Proceedings of 9th ICDAR, pp.709-713, 2007.

[34] S. Sinha, U. Pal, B.B. Chaudhuri, "Word–Wise Script Identification from Indian Documents", LNCS 3163, S. Marinai and A. Dengel (Editors), pp. 310–321, 2004.

[35] U. Pal, S. Sinha, B. B. Chaudhuri, ―Multi-Script Line identification from Indian Documents‖, in Proc. 7th Int. Conf. on Document Analysis and Recognition, pp.880-884, 2003.

[36] R.M.K. Sinha, ―Rule based contextual post-processing for Devnagari text recognition‖, Pattern Recognition, Vol.20, pp.475–485, 1987.

[37] K. Jayanthi, A. Suzuki, H. Kanai, Y. Kawasoe, M. Kimura, K. Kido, ―Devanagari Character Recognition Using Structure Analysis‖, in Proceedings of the IEEE-TENCON, pp.363-366, 1989.

[38] V. Govindaraju, S. Khedekar, S. Kompalli, F. Farooq, S. Setlur, R.Vemulapati, tools for enabling digital access to multilingual indic documents‖, In Proceedings of the 1st International Workshop on Document Image Analysis for Libraries, pp.122–133, 2004.

[39] S. Kompalli, S. Setlur, V. Govindaraju, ―Design and Comparison of Segmentation Driven and Recognition Driven Devanagari OCR‖, in Proceedings of the Second International Conference on Document Image Analysis for Libraries, pp.1-7, 2006.

[40] A. Dhurandhar, K. Shankarnarayanan, R. Jawale, ―Robust Pattern Recognition scheme for Devanagari Script‖, CIS, Part I, LNAI 3801, pp.1021 – 1026, 2005.

[41] M Meshesha, C V Jawahar, ―Matching word images for contentbased retrieval from printed document images‖, International Journal of Document Analysis and Recognition, Vol.11, pp.29–38, 2008.

[42] S. Chaudhury, G. Sethi, A. Vyas, G. Harit, "Devising Interactive Access Techniques for Indian Language Document Images", in Proceedings of 7th ICDAR, pp.885-889, 2003.

[43] P. Natarajan, E. MacRostie, M. Decerbo, "The BBN Byblos Hindi OCR System", Guide to OCR for Indic Scripts, V. Govindaraju and S. Setlur (Editors), Springer-Verlag, pp. 173-180, 2009.

[44]. M. Agrawal, H. Ma, D. Doermann, "Generalization of Hindi OCR Using Adaptive Segmentation and Font Files", Guide to OCR for Indic Scripts, V. Govindaraju and S. Setlur (Editors), Springer-Verlag, pp. 181-208, 2009.

[45] U. Pal, P. P. Roy, N. Tripathy , J. Llados, ―Multi-Oriented Bangla and Devnagari Text Recognition‖, Pattern Recognition, 2010

[46] V. Bansal, R. M. K. Sinha, ―A Complete OCR for Printed Hindi Text in Devanagari Script‖, in Proceedings of Sixth International Conference on Document Analysis and Recognition, pp.800-804, 2001.

[47] V. Bansal and R.M.K. Sinha, ―Partitioning and searching dictionary for correction of optically read Devanagari character strings‖, International Journal of Document Analysis and Research, Vol.4, pp.269–280, 2002.

[48] K. D. Dhingra, S. Sanyal, P. K. Sharma, ―A robust OCR for degraded documents‖, Advances in Communication Systems and Electrical Engineering, Huang et al., (eds.), Lecture Notes in Electrical Engineering, Springer, pp. 497-509, 2008.

[49] XuewenWang, Xiaoqing Ding and Changsong Liu, ―Gabor Based Feature extraction for Character Recognition‖, Pattern Recognition, Vol.38, pp.369–379, 2005

[50] T. Kanungo,R. M. Haralick, H. S. Baird, W. Stuezle, D. Madigan, ―A Statistical, Nonparametric Methodology for Document Degradation Model Validation‖ , IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol.20, pp.1209–1223, 2000.

[51] U. Pal, T. Wakabayashi, F Kimura, ―Comparative study of Devnagari handwritten character recognition using different features and classifiers‖, in Proceedings of 10th International Conference on Document Analysis and Recognition, pp.1111-1115, 2009.

[52] B. Shaw, S. K. Parui, M. Shridhar, ―Off-line Handwritten Devanagari Word Recognition: A holistic approach based on directional chain code feature and HMM‖, in Proceedings of International Conference on Information Technology, pp.203-208, 2008.

[53] N. K. Garg, L. Kaur, M. K. Jindal, "A New Method for Line Segmentation of Handwritten Hindi Text", in Proceedings of 7th ICIT, pp.392-397, 2010.

[54] A. Bhardwaj, S. Setlur, V. Govindaraju, "Keyword Spotting Techniques for Sanskrit Documents", LNAI 5402, G. Huet, A. Kulkarni, and P. Scharf (Editors), pp. 403–416, 2009.

[55] U. Pal, P. K. Kundu, B. B. Chaudhuri,"OCR error correction of an Inflectional Indian language using morphological parsing", Journal of Information Science and Engineering, Vol. 16. No. 6. pp. 903-922, 2000.

[56] A. Tarafdar, R. Mandal, S. Pal, U. Pal, Fumitaka Kimura, "Shape Code based Word-image Matching for Retrieval of Indian Multilingual Documents", in Proc. 20th International Conference on Pattern Recognition, 2010

[57] V. Vapnik, ―The Nature of Statistical Learning Theory‖, Springer Verlang, 1995.

[58] U. Pal, P. P. Roy, N. Tripathy , J. Llados, ―Multi-Oriented Bangla and Devnagari Text Recognition‖, Pattern Recognition, 2010

[59] Richard G. Cassey and Eric Lecolinet. A Survey of methods and strategies in character segmentation, EEE Transactions on Pattern analysis and Machine Intelligence, Vol 18, No. 7, 1996.

[60] U. Bhattacharya, S. Vajda, A. Mallick, B. B. Chaudhuri, A. Belaid, ―On the Choice of Training Set, Architecture and Combination Rule of Multiple MLP Classifiers for Multiresolution Recognition of Handwritten Characters‖, 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004).

[61] M. Egmont-Petersen, D. de Ridder, H. Handels, ―Image Processing with Neural Networks: A Review‖, Pattern Recognition, Vol 35, pp. 2279–2301, 2002

[62] K. Y. Rajput and Sangeeta Mishra, ―Recognition and Editing of Devnagari Handwriting Using Neural Network‖, SPIT-IEEE Colloquium and Intl. Conference, Mumbai, India.